JOURNAL of COMPUTER SCIENCES INSTITUTE

JCSI 8 (2018) 230-234
Wysłane: 2017-12-12
Przyjęte: 2017-12-19

System monitoringu użytkownika wykorzystujący sieci społecznościowe – budowa i analiza możliwości

Sofiia Lahoda *, Marek Miłosz
Politechnika Lubelska, Instytut Informatyki, Nadbystrzycka 36B, 20-618 Lublin, Polska

Streszczenie. W artykule przedstawiono rezultaty badań metod, które pomogą tworzyć bazę danych o użytkownikach z wykorzystaniem serwisów internetowych jako podstawowego źródła informacji. Celem badania było pozyskanie danych o użytkownikach z sieci społecznościowej, sprawdzenie możliwości parsera oraz analiza efektywności wykorzystania utworzonej bazy danych. W badaniach zostały wykorzystane metody analizy tekstowej: parsing i scraping. Wyniki zostały przedstawione w postaci wykresów i poddane krytycznej analizie porównawczej.

Słowa kluczowe: parsowanie; scraping; baza danych; sieci społecznościowe; analiza tekstowa

*Autor do korespondencji.

Adres e-mail: lahoda.sof@gmail.com

User monitoring system using of social networks - structure and analysis of the opportunities

Sofiia Lahoda *, Marek Miłosz
Institute of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract. The article presents the results of a study that will help for user to create a end-user parameters database using other web-sites as the primary source. The main goal of the work is making analysis to obtain information about the user of social networks. Next goal is analysis of the gathering data and the possibility of their use. The experiment was done using two methods of text analysis: parsing and scraping. The results are presented graphically and critical compared to each other.

Keywords: parsing; scraping; data bases; social networking; text analysis

*Corresponding author.

E-mail address: lahoda.sof@gmail.com

1. Wstęp

W obecnej sytuacji, oprogramowanie, które jest niezbędne do wyszukiwania i analizowania informacji w obszarze sieci społecznościowych jest niezbędnym narzędziem dla działów analitycznych wielu firm w kraju - od małych przedsiębiorstw do dużych korporacji. Wyłączając bowiem ze swojej działalności użytkowników portali społecznościowych, takich jak "Facebook", "Instagram" czy "Twitter", firma traci wielu potencjalnych klientów, a co za tym idzie, traci potencjalne zyski. Walcząc o rynek firmy powinny walczyć także o członków sieci społecznościowych.

Informacje, które użytkownicy udostępniają w sieciach społecznościowych, najczęściej zawierają dane, w których zawarte są opinie na temat marki, produktu czy usługi. Używając danych z sieci społecznościowych, użytkownik otrzymuje możliwość wyszukiwania interesujących go informacji, a także zajmuje się aktualizacją danych, dokonuje selektywnego monitorowania interesującej informacji i jej pełnej analizy.

Często producent i jego menedżer nie wie, w którym kierunku ma iść, jak działać, aby osiągnąć swoje cele, czego potrzebuje klient, z którego miasta jest najwięcej kupujących itp.

Aby rozwiązać podobne problemy, konieczne jest systematyczne prowadzenie badań mających na celu analizę obiektywnej sytuacji społecznej i gospodarczej w danym obszarze tematycznym lub po prostu badań statystycznych dla np. miasta czy kraju.

Na wymagania takiej analizy całkowicie odpowiada monitoring społeczny jako metoda monitoringu sieci społecznościowych. Dużą zaletą monitoringu, w porównaniu z jednorazowymi badaniami, jest jego zdolność do systematycznego zwiększania i integrowania niezbędnych danych w szerokim zakresie standardowych wskaźników społecznych i tworzenia na tej podstawie stale aktualizowanych zbiorów danych. Pozyskiwanie danych jest permanentne i właśnie dlatego rośnie popularność tej metody w praktyce analizy socjologicznej.

Jednym z narzędzi, za pomocą których można przeprowadzić takie monitorowanie, są wspomniane wcześniej sieci społecznościowe. Za ich pomocą można uzyskać różnorodne informacje do analizy. Monitorowanie

obejmuje badanie faktów, zdarzeń i wyników związanych z przedmiotem obserwacji. Taką analizę sieci społecznościowej można wykonać za pomocą specjalnych aplikacji, bazujących na użyciu metod analizy składniowej danych.

2. Cel, teza i zakres artykułu

Celem artykułu jest zaprezentowanie rezultatów badań o metodach pozyskania informacji o użytkowniku z sieci społecznościowych do ich wykorzystania w różnych obszarach. Ocenione będą także korzyści z wykorzystania danych monitorowania z sieci społecznościowych i potencjalne obszary ich zastosowania.

Dla celów badań sformułowano następujące hipotezy:

- Hipoteza 1: Przy pomocy parsowania jest możliwe monitorowanie użytkowników sieci społecznościowej.
- Hipoteza 2: Monitorowanie użytkowników w sieci społecznościowej pomaga stworzyć bazę danych o użytkownikach sieci społecznościowej, którą można wykorzystać w różnych obszarach.

Tezą danej pracy jest: Monitoring użytkowników sieci społecznościowej pomaga uzyskać informacje, które nadają się do analizy biznesowej i statystycznej.

Zakres artykułu obejmuje:

- badanie literatury na temat analizy danych;
- analizowanie pracy parserów – analizatorów składniowych;
- analiza informacji na temat tworzenia parsera w języku programowania php;
- opracowanie i opis działania stworzonego programu do monitorowania użytkowników sieci społecznościowej;
- użycie stworzonego programu oraz pozyskanie i rejestrowanie danych użytkownika w bazie danych;
- przeprowadzenie eksperymentu na dwóch grupach w celu porównania skuteczności dostępności danych uzyskanych z sieci społecznościowych;
- wnioski.

3. Analiza składniowy danych

Analiza składniowa (ang. parsing) - jest sprawdzaniem poprawności składniowej programu lub modułu wykonywanym przez translator (kompilator) i polega na dokonaniu rozbioru gramatycznego analizowanej jednostki tekstu [1].

Zadanie analizy składniowej można rozpatrywać jako potwierdzenie lub zaprzeczenie zgodności tekstu na wejściu z daną gramatyką formalną, a w przypadku stwierdzenia zgodności — także wygenerowanie wszystkich drzew możliwych rozbiorów gramatycznych tekstu. Ogólnie takich rozbiorów może okazać się dużo, nie tylko z powodu znaczących różnic w strukturyzacji, ale także z powodu wielu możliwości grupowania struktur [2]. Z tej przyczyny, nawet jeśli zadanie analizy składniowej może być zrobione ręcznie, stopień złożoności języka naturalnego znajdujący odzwierciedlenie w złożoności opisującej jego gramatyki sprawia, że tylko analiza automatyczna ujawnia większą część potencjalnych niejednoznaczności rozbioru.

Analizator składniowy lub parser jest narzędziem, które służy do analizy przetwarzanych danych oraz ewentualnie wywołuje określone akcje. Analizator leksykalny lub skaner analizuje znaki i przekazuje te do parsera w postaci tokenów. Istnieje dużo narzędzi do tworzenia takich analizatorów składniowych oraz leksykalnych.

Drzewo rozbioru jest jedną z podstaw parsowania. Wierzchołki drzewa rozbioru oznacza [3]: albo symbolem terminalnym, albo symbolem ε, albo kategoriami syntaktycznymi. Tak zwane liście etykietuje się jedynie symbolami terminalnymi albo symbolem ε. Wierzchołki wewnętrzne są etykietowane tylko kategoriami syntaktycznymi. Każdy wewnętrzny wierzchołek reprezentuje zastosowanie otrzymanych danych.

Z tego wynika, że kategoria syntaktyczna, która etykietuje wierzchołek stanowi część nagłówka danych. Każde drzewo rozbioru proponuje ciąg symboli terminalnych s, który można nazwać wynikiem drzewa [3].

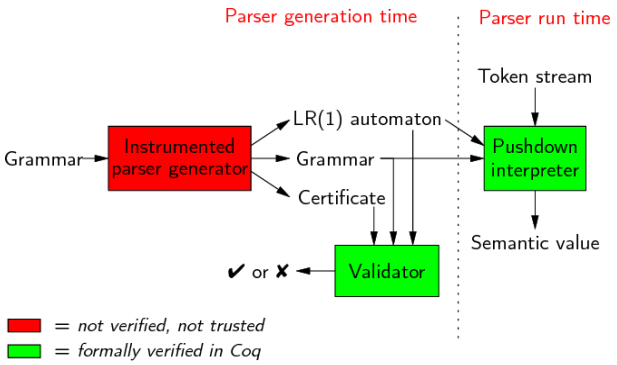
Ciąg ten składa się z tak zwanych etykiet liści drzewa ułożonych w specjalnej kolejności, od lewej do prawej strony. W przypadku, gdy drzewo posiada tylko jeden wierzchołek, on jest etykietowany symbolem terminalnym lub symbolem ε, ponieważ on jest liściem.

Gdy drzewo posiada więcej niż jeden wierzchołek, korzeń etykietuje się kategorią syntaksyczną, dlatego że korzeń drzewa, który posiada dwa lub więcej wierzchołków, zawsze będzie wierzchołkiem wewnętrznym. W tej kategorii syntaktycznej zawsze będzie wśród ciągów znaków również wynik drzewa.

Zazwyczaj wynikiem analizy składniowej będzie struktura składniowa zdania, zaproponowana albo w postaci drzewa komponentów, albo jako drzewo zależności, lub jako kombinacja pierwszej i drugiej przedstawianej metody [4].

Jakie są typy algorytmów do tworzenia parserów? Nie ma ich zbyt dużo. Pierwszy to rekurencyjny analizator syntaksyczny, który analizę zaczyna od znaku początkowego i kontynuuje do otrzymania wymaganej liczby tokenów [5]. On odpowiada za parser LL (Leftmost derivation). Drugim jest parser oddolny - zaczyna odzyskiwane z prawej części, zaczynając od tokenów i kończy na symbolu startowym. On dzieli się na parser LR (Rightmost derivation) (przykład na rysunku 1) oraz parser GLR (Generalized Left-to-right Rightmost derivation parser).

Parsowanie stron jest oczywiście analizą danych publikowanych na stronach internetowych. Tekst stron internetowych jest hierarchicznym zbiorem danych, który funkcjonuje razem z językiem, na przykład, angielskim i komputerowym. Język ludzki oferuje informacje, wiedzę. Języki komputerowe, na przykład - html, JavaScript, css - określają sposób wyświetlania informacji na monitorze [7].



Rys. 1. Proces parsera LR [6]

4. Eksperyment

W czasach współczesnych bardzo popularnym wśród programistów jest połączenie PHP i MySQL. PHP jest językiem programowania dość łatwym w użyciu, a jednocześnie bardzo elatycznym. Przy jego pomocy można także tworzyć połączenie z bazą danych MySQL. Pomaga on deweloperom tworzyć dynamiczne elementy na stronach.

W związku z powyższym, do napisania własnego parsera był wybrany język PHP. Spośród innych wariantów on idealnie odpowiadał wszystkim potrzebom.

Każda strona internetowa składa się z kodu HTML z dodatkowymi elementami technologicznymi (JavaScript, PHP, CSS i inne). Zadanie pozyskania informacji w tym przypadku polega na znalezieniu i wyodrębnieniu wzorców tekstowych, czyli sekwencji istniejących symboli, które spełniają zasady podane przez gramatykę [8].

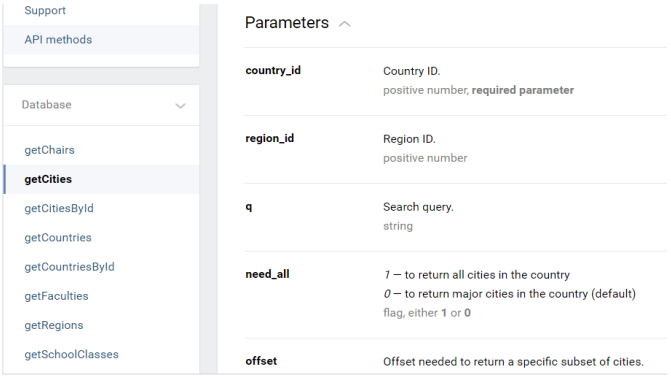
Parsowanie strony HTML to proces, który można podzielić na trzy etapy:

1) Uzyskanie kodu źródłowego z potrzebnej strony. W tym celu w różnych językach dostępne są odpowiednie metody. Na przykład PHP używa biblioteki cURL lub wbudowanej funkcji, takiej jak file_get_contents [9].

2) Otrzymanie niezbędnych danych z kodu HTML. Po otrzymaniu strony należy ją przetworzyć, w tym przypadku - oddzielić zwykły tekst od języka znaczników hipertekstowych, zbudować hierarchiczne drzewo struktury dokumentu, poprawnie zareagować na nieprawidłowy kod, wyodrębnić z tej strony dokładnie te informacje, dla pozyskania których wykonano całą aplikacje. Oczywiście można również użyć wyrażeń regularnych, ale łatwiejszym sposobem jest używanie biblioteki, która w tym specjalizuje się.

3) Zapisywanie wyniku. Po bezpiecznym przetworzeniu danych ze strony, trzeba zapisać je w określonym formacie. Otrzymane dane po parsowaniu są zwykle zapisywane do bazy danych, ale oprócz tego możliwe są i inne opcje. Czasami można zapisywać na przykład do pliku CSV (ang. Comma-Separated Values) lub w postaci hierarchicznej struktury JSON (ang. JavaScript Object Notation).

Dla napisania własnego parsera trzeba wykonać powyższe kroki. Parsowanie w artykule było wykonane dla sieci społecznościowej “Vkontakte”. Stworzona aplikacja pobierała dane użytkowników sieci społecznościowej i zapisywała je do bazy danych. Do uzyskania danych było użyte API sieci “Vkontakte”, które pomaga filtrować użytkowników odpowiednio wg miejsca zamieszkania, płci, wieku i innych kryteriów. Do pobierania danych była użyta sieciowa biblioteka cURL, także za jej pomocą dane były oddzielone od całego kodu HTML. Pozyskane dane są przypisywane do odpowiednich kategorii i automatycznie zapamiętywane w bazie danych. Aplikacja ma możliwość ściągać dane osób, które mieszkają wyłącznie w Warszawie. Dodatkowa funkcja to pobieranie danych z grup. Otrzymane dane można wykorzystać, na przykład, w reklamie [10]. Po użyciu metody GroupSearch API sieci „Vkontaktie” nie ma potrzeby pisać własnej funkcji, która będzie odpowiadać za ten proces (przykład na rysunku 2). Oprócz tego, aplikacja ma możliwość automatycznie publikować wiadomości do potrzebnych grup. Taką możliwość nadaje skrypt access_token.



Rys. 2. Metody wykorzystane z API sieci społecznościowej „Vkontaktie”

5. Rezultaty badań

Po napisaniu aplikacji, można użyć jej funkcjonalności i zweryfikować postawione hipotezy. Czy, jak było umówione wyżej, jest możliwość pozyskać dane użytkowników z sieci społecznościowej, i czy są z tego korzyści?

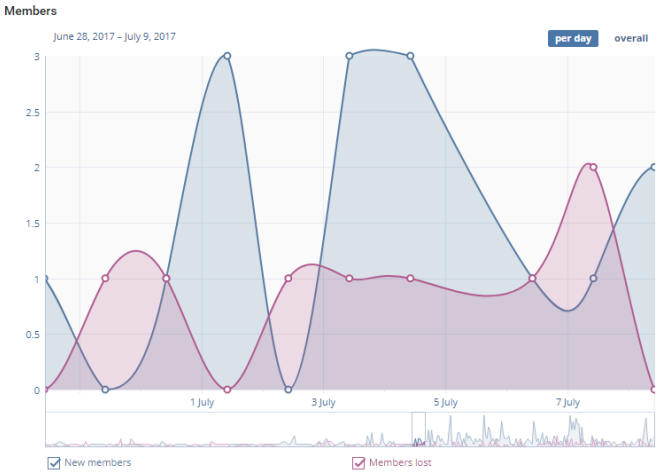
Na początku badania stworzone były dwie grupy użytkowników. Główny element różnicującym grupy są parametry członków zapraszanych do obydwu grup. Uczestnicy pierwszej grupy byli wybrani losowo. Dla określenia użytkowników zapraszanych do drugiej grupy był użyty parser. Aplikacja wyszukała użytkowników i ze wszystkich użytkowników, wybrała wyłącznie mieszkańców Warszawy, pobrała ich dane i zapisała do tabeli. Otrzymane dane były pobierane w postaci XML i zawierały różne pola jak pokazano na rysunku 3.

Analogicznie była stworzona baza danych użytkowników, posiadających wpisy ze wspólną tematyką. Wszystkie otrzymane dane użytkowników były zapisane do bazy danych. Każdy z nich był zaproszony do grupy społecznościowej. Dzięki danym statystycznym,

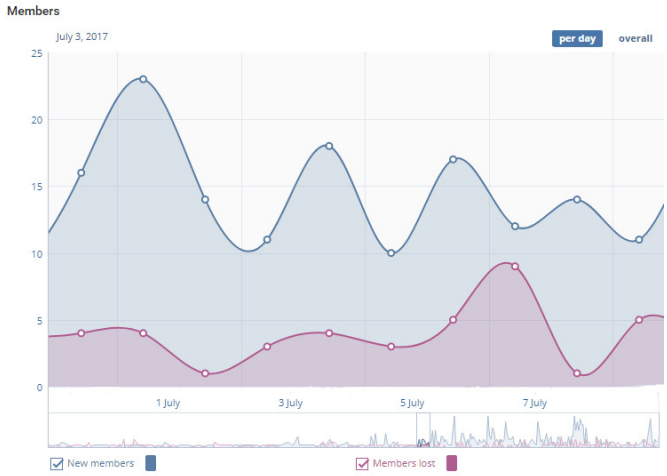
udostępnianym w wybranej sieci społecznościowej, można było porównać postęp rozwoju dwóch grup.

	A	B	C	D	E	F	G	H
1	Nazwisko	Imie	Link	Płeć	Kraj	Miasto	Data urod	Wiek
2	Zheka	Kocherov	http://vk.	m	Polska	Warszawa		
3	Irina	Porada	http://vk.	k	Polska	Warszawa	########	31
4	Sanya	Kitayev	http://vk.	m	Polska	Warszawa	08.янв	0
5	Maxim	Chornoba	http://vk.	m	Polska	Warszawa		
6	Mikhail	Baytsar	http://vk.	m	Polska	Warszawa	########	48
7	Vadim	Muzichuk	http://vk.	m	Polska	Warszawa	06.мар	0
8	Vitaly	Easy-Goin	http://vk.	m	Polska	Warszawa	19.дек	0

Rys. 3. Baza danych użytkowników cieci „Vkontakte” z Warszawy



Rys. 4. Wykres dodawania użytkowników do grupy A

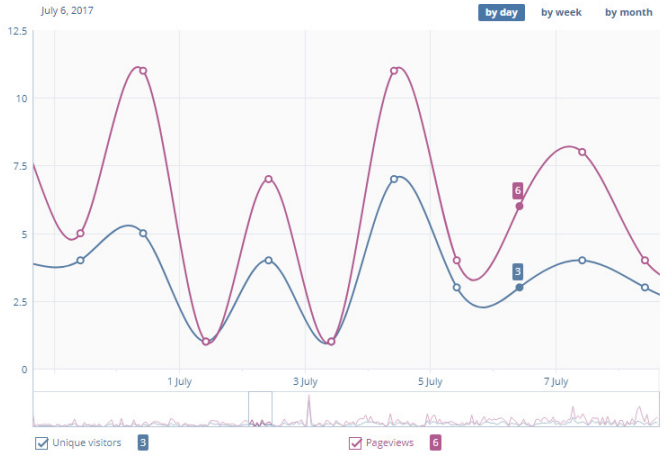


Rys. 5. Wykres dodawania użytkowników do grupy B

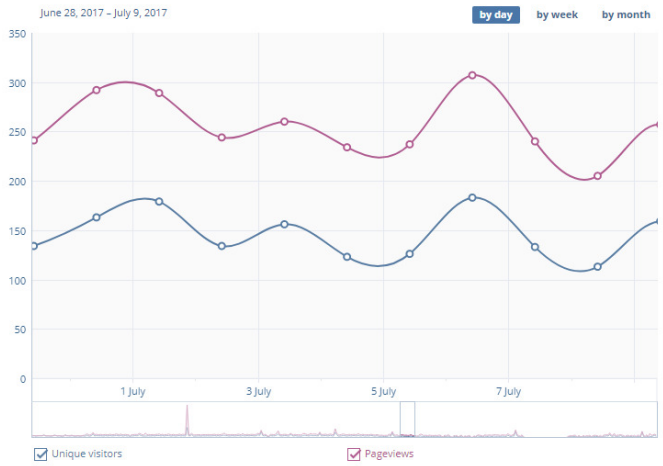
Na rysunkach 4 i 5 jest pokazana liczba nowych użytkowników obydwu grup. Dziennie do grupy, która nie była opracowana parserem tj. do grupy A, średnio dołączało się około 3 osób dziennie. Do grupy B, użytkownikami której były zainteresowane osoby, dołączało się koło 15 osób dziennie. Z takich wykresów można wywnioskować, że dołączenie do grupy osób wyselekcjonowanych na podstawie miasta zamieszkania i zainteresowań jest bardziej efektywne. Taki monitoring osób przez sieć społecznościową daje administratorowi możliwość odszukanie bardziej zainteresowanych danym tematem (w tym przypadku zainteresowań i miejsca) członków sieci społecznościowej.

Wiadomo, że duża liczba użytkowników to jeszcze nie sukces. Dlatego sprawdzono na ile użytkownicy są zainteresowani. Na rysunku 6 znajduje się wykres z pierwszej grupy A, która nie zainteresowała użytkowników. Oglądalność strony jest bardzo niska, użytkownicy nie są zainteresowani tym, co publikuje się w grupie. Wykres 6 pokazuje statystykę, otrzymaną po miesiącu utworzenia grupy przy 30 członkach grupy.

Przy porównaniu wykresów 6 i 7 można powiedzieć, że liczba wejść na stronę dwóch grup dużo różni się. W tym przypadku, liczba ta w grupie A równa się 5, a w grupie B, która była przygotowana za pomocą parsera – średnio 300 dziennie.



Rys. 6. Wykres wejść użytkowników do grupy A



Rys. 7. Oglądania wykres wejść użytkowników do grupy B

6. Wnioski

Celem danego artykuła było sprawdzanie hipotezy, czy monitorowanie użytkowników w sieci społecznościowej pomaga stworzyć bazę danych, która zawiera dane o wyselekcjonowanych użytkownikach sieci

społecznościowej do wykorzystania w różnych obszarach. Hipoteza 1 została zatem całkowicie udowodniona.

Hipoteza 2 również jest prawdziwa. Osoby wyselekcjonowane na podstawie danych pozyskanych z sieci społecznościowej były bardziej aktywne (i zainteresowane tematem) niźli te przypadkowe.

Wykorzystując informacje pozyskane ze stron członków grupy społecznościowej, można stworzyć bazę danych klientów ukierunkowanych na określony cel. Wynika to z faktu, że pozyskać można nie tylko dane o użytkowniku, ale też o jego zainteresowaniach.

Trzeba pamiętać, że dane mogą być zastrzeżone przez stronę internetową, dlatego trzeba sprawdzić reguły i wymagania strony przed jej parsowaniem.

Tworzenie parsera nie jest zbyt ciężkim zadaniem dla programisty, ponieważ istnieje dużo bibliotek oraz API, które ułatwiają ten proces.

Istnieje duża liczba parserów, których użycie i wykorzystanie rezultatów ich pracy umożliwia wyselekcjonowanie grupy docelowej w działaniach marketingowych.

Hipotezy zostały udowodnione i potwierdzone przy pomocy własnej aplikacji.

Literatura

[1] M. Collins, J. Ha, E. Brill, L. Ramshaw, C. Tillmann, A Statistical Parser for Czech. In Proceedings of ACL, University of Maryland, College Park, USA (1999), 505-512.

[2] M. Collins, Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania, 1999.

[3] https://www.parser.ru/ [28.04.2016]

[4] Hausser R.: Computation of Language, Springer-Verlag, 1989.

[5] http://myblaze.ru/chto-takoe-parser-grabber/ [4.01.2015]

[6] http://gallium.inria.fr/blog/verifying-a-parser-for-a-c-compiler/ [24.10.2012]

[7] Srinivas B., Doran C., Hockey B., Sarkar A.: Gramar & Parser Evaluation in the XTAG Project, Proceedings of the 1st International Conference on Language Resources and Evaluation. Granda, 1998.

[8] Daniel D. K., Temperley D., Parsing English with a Link Grammar, 1991.

[9] Stump G.T.: Morphological and syntactic paradigms: arguments for a theory of paradigm linkage, Yearbook of Morphology (2001), 147-80.

[10] http://excelvba.ru/programmes/Parser [14.11.2017]